EXPLORATION OF A BIOLOGICALLY INSPIRED MODEL FOR SOUND SOURCE LOCALIZATION IN 3D SPACE

Symeon Mattes, *

ISVR Acoustics Group
University of Southampton,
Southampton, UK
symeon.mattes@soton.ac.uk

Philip Arthur Nelson

ISVR Acoustics Group
University of Southampton,
Southampton, UK
p.a.nelson@soton.ac.uk

Filippo Maria Fazi

ISVR Acoustics Group
University of Southampton,
Southampton, UK
ff1@isvr.soton.ac.uk

Michael Capp

Research Department
Meridian Audio Ltd,
Cambridgeshire, UK
Michael.Capp@meridian.co.uk

ABSTRACT

Sound localization in 3D space relies on a variety of auditory cues resulting from the encoding provided by the lower and higher regions of the auditory path. During the last 50 years different theories and models have been developed to describe psychoacoustic phenomena in sound localization inspired by the processing that is undertaken in the human auditory system. In this paper, a biologically inspired model of human sound localization is described and the encoding of the known auditory cues by the model is explored. In particular, the model takes as an input binaural and monaural stationary signals that carry information about the Interaural Time Difference (ITD), the Interaural Level Difference (ILD) and the spectral variation of the Head Related Transfer Function (HRTF). The model processes these cues through a series of linear and nonlinear units, that simulate the peripheral and the pre-processing stages of the auditory system. The encoded cues, which in the model are represented by excitation-inhibition (EI) and the time average (TA) activity patterns, are then decoded by a central processing unit to estimate the final location of the sound source.

1. INTRODUCTION

Sound localization is a perceptual process that in contrast to other sensory systems, like vision and taste, there is no point-to-point correspondence between a sound event and the perceived locus of an acoustic image at the lower peripheral stages of the human hearing system [1]. Instead, it is believed that the localization of sound events occur entirely as a consequence of neural processing of monaural and binaural signals. The ITDs (interaural time differences), the ILDs (interaural level differences), and the monaural spectral cues, that occur due to the spectral changes of the pinna, are three of the most salient auditory cues that are used by a human listener in order to characterize the locus of a sound event.

During the last 50 years different techniques have been developed to predict the statistical properties of human sound localization in the horizontal plane. Some of these theories rely only on stimulus statistics, while others are based on neuroscientific findings. The last one has led to the development of so called biologically inspired models and to three of the most established and well-known theories, i.e. the Jeffress's coincidence detector [2], that is based on coincidence counter hypothesis, Durlach's EC (equalization-cancellation) theory [3], that was developed to interpret phenomena in the detection of binaural sounds masked by a masking noise, and the count-comparison principle introduced by von Békésy (1930) [4] that resembles the neural activity of the higher regions of the auditory path.

At the same time only recently, a variety of different models have been developed for the prediction of human sound localization in sagittal planes [5, 6]. These models are based mainly on the neural integration hypothesis, which states that for moderate intensities the localization system requires an input of at least 80 ms broadband sound to give a stable estimation of the sound-source elevation [7, 8].

Having such models, i.e. a model that is able to predict successfully under certain conditions, human modes of listening, can be beneficial not only for the better understanding of the underlying mechanisms of human reactions but also for their application in audio quality assessment, robotics and cochlear implants, avoiding costly and time-consuming experiments.

The current paper aims to combine two well established models for the prediction of human localization in horizontal and sagittal planes in order to predict human localization in 3D space. The paper is divided into five main sections. In the first section a general introduction to sound localization and to perceptual models is given and in the second section, a biologically inspired model is described for the prediction of human sound localization for stationary signals in 3D space (excluding distance). In the third section, different parameters of the model are explored, and in the third section, simulation results are compared with previous listening tests. In the last part the conclusions and future work are given.

2. DESCRIPTION OF THE MODEL

The model that has been used in this paper is based on EC theory for the production of the excitation-inhibition (EI) pattern in binaural processing [9], which is mainly responsible for the encoding of the ITD and ILD cues, and a time average (TA) representation

* This work was supported by Meridian Audio Ltd.

of a narrow band filtered signal for the production of the monaural processing [6], which is responsible for the encoding of the spectral variations of the HRTFs.

In particular, the model consists of three main stages, each of which corresponds to different (and more or less known) operations of the human auditory system in spatial hearing. The model starts with the peripheral processor, which takes binaural signals as an input. This stage consists of a unit which corresponds to a time-invariant band pass filter from 1 kHz - 4 kHz with a roll-off of 6 dB/octave below 1 kHz and -6 dB/octave above 4 kHz, which represents the response of the human middle ear. This is followed by a fourth-order gammatone filterbank with 100 channels between 100 Hz and 20 kHz [10], which represent the frequency selectivity of the basilar membrane. Each gammatone filter output is processed by a half-wave rectifier, a fifth-order low pass filter with a cut-off frequency at 770 Hz, and a square root compressor, which respectively represents the organ of Corti [11], the gradual loss of the phase-locking in neural firing [12], and the nonlinearities of the basilar membrane in steady state conditions[13].

The model continues with the pre-processor, which consists of one binaural and two monaural units. Each of these units creates three types of patterns ($EI_{k,\tau,\alpha}$, TA_{L_k} and TA_{R_k}) correspondingly, that are compared in the central-processor with a database of patterns by applying a comparison metric which consists of frequency independent functions (m_{bin}, m_L and m_R), called similarity measure (SM) functions [14]. A mapping function is applied to transform m_{bin}, m_L and m_R into the transformed similarity measure function s_{bin}, s_L and s_R. All these functions are then combined to give a single function that represents the likelihood of subject localization of the virtual source.

More specifically, in the pre-processor, the binaural unit, as described by Park et al. [9], is based on the EC theory for the extraction of the excitation-inhibition (EI) cell activity patterns (EI-patterns) and is responsible for the characterization of the position of a lateralized sound source. Given that $L_k(t)$ and $R_k(t)$ are the input signals from the left and the right peripheral processor from the k-th channel of the gammatone filterbank, then each EI unit is characterized by the equation

$$EI_{k,\tau,\alpha}(t) = \left(10^{\frac{\alpha}{40}} L_k(t+\frac{\tau}{2}) - 10^{-\frac{\alpha}{40}} R_k(t-\frac{\tau}{2})\right)^2 \quad (1)$$

where τ is the characteristic ITD in seconds and α the characteristic ILD in dB that occur due to the comparison of the signals of the left and the right ear. At 44.1 kHz sampling frequency the dynamic range is $\pm 700\mu$sec for the characteristic ITD and ± 10 dB for the characteristic ILD, with a resolution of 45 μsec and 1 dB respectively.

Thereafter, the EI-cell activity is normalised by the energy of the input signals associated with a specific snapshot in time, so as to remove any dependency of the amplitude of the input signal. In this case the binaural unit is described by the equation

$$EI''_{k,\tau,\alpha} = \frac{EI'_{k,\tau,\alpha}}{\sqrt{2 e_L e_R}} \quad (2)$$

where e_L and e_R are the energy of the left and the right input signals correspondingly and $EI'_{k,\tau,\alpha}$ is an integrated weighted snapshot over the time t, defined as

$$EI'_{k,\tau,\alpha}(t) = \int EI_{k,\tau,\alpha}(t+t')w(t')\mathrm{d}t' \quad (3)$$

and $w(t)$ is a double-sided exponential window that takes into account the finite binaural temporal nature of the EI-cell activity [9].

The two monaural units are based on the hypothesis that a time average (TA) representation of the narrow band filtered signal that arrives from the peripheral processing unit can be used for the representation of the spectral variations that are necessary for the characterization of an elevated sound source. In this case each unit is characterized by the equation

$$y_k(t) = \frac{1}{T}\int_0^T x_k(t)\mathrm{d}t \quad (4)$$

where $x_k(t)$ is the output of each of the k gammatone filters for the left ($L_k(t)$) and the right ear ($R_k(t)$) integrated over a snapshot of the signal of duration T, which for the current paper the whole duration of the signal has been taken, and $y_k(t)$ is the corresponding monaural pattern for the left (TA_{L_k}) and the right (TA_{R_k}) ear.

The model ends with the central processing unit which is a decision making device that uses a simple pattern matching process in order to characterize the location of the sound source in 3D space. More specifically, the EI-patterns and the TA-patterns that have been produced by a sound source from an unknown location are compared with a bank of EI- and TA-pattern templates in order to produce a SM that quantifies the degree to which the patterns produced by a given source matches the stored patterns.

Given the stationarity and the uniqueness of the sound source, a pattern-matching procedure has been applied for measuring the similarity of the EI-patterns at each channel k of the gammatone filterbank and is defined as

$$\rho_{bin_k}(\phi,\theta) = \frac{\langle EI''_{k,\tau,\alpha},\, EI''_{k,\tau,\alpha}(\phi,\theta)\rangle}{\|EI''_{k,\tau,\alpha}\|\|EI''_{k,\tau,\alpha}(\phi,\theta)\|} \quad (5)$$

where ϕ and θ are the azimuth and elevation angle of the sound source in the interaural-polar coordinate system (fig. 1), $EI''_{k,\tau,\alpha}$ is the EI-patterns of eq. 2 of the target source for a specific azimuth ($\hat{\phi}$) and elevation angle ($\hat{\theta}$), $EI''_{k,\tau,\alpha}(\phi,\theta)$ is the template of the EI-patterns of eq. 2 for all possible azimuth (ϕ) and elevation (θ) positions at the same snapshot, $\langle\cdot\rangle$ is the inner product and $\|\cdot\|$ is the L^2 norm of the EI'' over τ and α.

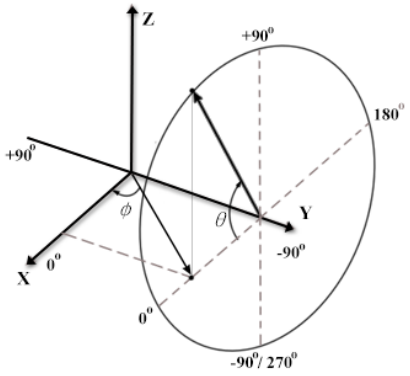

Figure 1: The interaural-polar coordinate system is a head-related spherical coordinate system whereby different azimuth angles ϕ define a cone of confusion. Its range for the azimuth angle is $\phi \in [-\frac{\pi}{2}, \frac{\pi}{2}]$ and for the elevation angle $\theta \in [-\pi, \pi)$ or $\theta \in [-\frac{\pi}{2}, \frac{3\pi}{2})$ [15, 16].

The frequency dependent SM is then weighted in order to give the total SM for the binaural cues, defined as

$$m_{bin}(\phi,\theta) = \sum_k \rho_{bin_k}(\phi,\theta) q_k \tag{6}$$

where q_k is a weighting coefficient that depends on the frequency of the gammatone filter and which varies smoothly with frequency but which reflects the dominance of the binaural cues around 600 Hz [17].

Finally, a mapping function is applied which gives the transformed SM for the binaural cues, defined as

$$s_{bin}(\phi,\theta) = m_{bin}(\phi,\theta)^{\gamma_{bin}} \tag{7}$$

where γ_{bin} modifies the transformed SM, and as demonstrated in sections 3.3 and 4, this will allow comparison with experimental data.

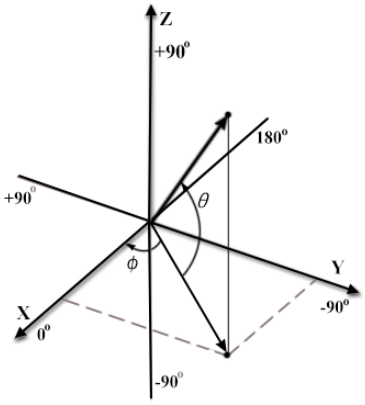

Figure 2: The vertical-polar coordinate system is a head-related coordinate system which is a sub-category of the spherical coordinate system. Its range for the azimuth angle is $\phi \in [-\pi, \pi)$ and for the elevation angle $\theta \in [-\frac{\pi}{2}, \frac{\pi}{2})$ [15].

The SM that has been used for the monaural cues is that suggested by Baumgartner et al. [6] and is the standard deviation of the interspectral differences, defined for the left monaural processor as

$$m_L(\phi,\theta) = \sqrt{\frac{1}{N}\sum_k \left(d_{L_k}(\phi,\theta) - \overline{d_{L_k}(\phi,\theta)}\right)^2} \tag{8}$$

where N is the number of the gammatone filters that has been used in the peripheral processing units, $d_{L_k}(\phi,\theta) = TA_{L_k} - TA_{L_k}(\phi,\theta)$ is the interspectral difference between the TA patterns (TA_{L_k}) of the target source (eq. 4) for a specific azimuth ($\hat{\phi}$) and elevation angle ($\hat{\theta}$) and the template of the TA-patterns ($TA_{L_k}(\phi,\theta)$) of eq. 4 for all available positions in the interaural coordinate system, and $\overline{d_{L_k}(\phi,\theta)}$ is the average value. Similar to eq. 8 , $m_R(\phi,\theta)$ gives the SM for the right monaural preprocessing unit.

Furthermore the SM of the monaural cues are combined through a weighted function as described by

$$s_{mon}(\phi,\theta) = b(\phi)s_L(\phi,\theta) + b(-\phi)s_R(\phi,\theta) \tag{9}$$

where $b(\phi)$ is a weighting function that is based on the assumption that the contralateral ear contributes less to the perception of sound localization than the ipsilateral ear [18], and

$$s_{L/R}(\phi,\theta) = \frac{1}{\sigma_{mon}\sqrt{2\pi}} \mathrm{e}^{-\frac{m_{L/R}(\phi,\theta)}{2\sigma^2_{mon}}} \tag{10}$$

is the mapping function, where $m_{L/R}(\phi,\theta)$ is the SM of the monaural cues for the left ($m_L(\phi,\theta)$) and the right ear ($m_R(\phi,\theta)$), and σ_{mon} again, as shown in sections 3.3 and 4, modifies the mapping function in a way that will allow comparison of the likelihood of localisation with experimental results.

By analogy with the laws of probability we multiply the two transformed SM ($s_{mon}(\phi,\theta)$ and $s_{bin}(\phi,\theta)$), as described by

$$s(\phi,\theta) = s_{bin}(\phi,\theta)s_{mon}(\phi,\theta) \tag{11}$$

to obtain a representation of the likelihood of the subject's localization of the virtual source.

3. EXPLORING THE LOCALISATION CUES

Two of the main characteristics of the model described in sec. 2 are the TA and EI patterns that are constructed through a process that attempts to emulate the human auditory path. These patterns contain information of the static cues associated with the ITD, the ILD and the spectral variations induced by the two pinnae and as a consequence information on the location of a given sound source. The aim of the following sections is to analyze some of the features of the TA and the EI patterns by using the HRTFs of a KEMAR with a small pinna from the CIPIC database (subject 165) [16].

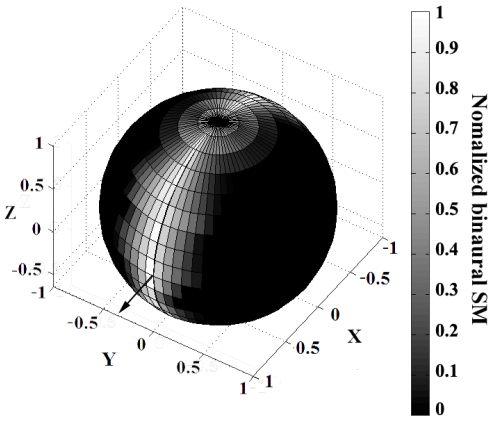

Figure 3: The results of the comparison of the EI patterns at $f = 100$ Hz for a sound source at $\hat{\phi} = \hat{\theta} = 0°$ by using the vertical-polar coordinate system. The colour bar indicates the value of eq. 5 normalised by its maximum value.

3.1. Binaural cues

The localization ambiguity arising from the cone of confusion can be resolved quite readily by head motion [1]. However, even if the head is restrained, partial resolution is still possible on the basis of the static spectral cues [19]. Resolution of the ambiguity is further improved if the listener has a priori information which restricts the possible source locations. For example, if the subject knows in advance that the sound source is in the horizontal plane in front. Considering these factors, it was necessary to verify the ability of the binaural unit of the model to resolve any static cues

of elevation, i.e. whether the EI-patterns are able to give any information of the location of an elevated source given that they only characterize the ITD and ILD cues.

Considering that the EI patterns depend on the frequency of the gammatone filterbank channel (k) of the peripheral processing unit, the azimuth ($\hat{\phi}$) and elevation angle ($\hat{\theta}$) of a target sound source, and the ITD (τ) and the ILD (α) that occurs due to the comparison of the signals of the left and the right ear, we compared the EI patterns created by a given $\hat{\phi}$ and $\hat{\theta}$ with all the EI patterns for all possible ϕ and θ in 3D space by using eq. 5.

In Figure 3 there are some representative results of the comparison of the EI patterns created by a white noise sound source at a given location $\hat{\phi}$, $\hat{\theta}$ in the vertical-polar coordinate system (fig. 2). From visual observation we can see that at low frequencies a clear circle is formed, which indicates a cone of confusion, and as a consequence, the inability of EI patterns to predict the location of elevated sources. Similar results have been obtained for frequencies up to 4kHz. This indicates that in low and middle range frequencies where the ITD cues are prominent, the EI patterns are not able to predict the location of elevated sources, however they give a clear indication of the lateralized sources. At higher frequencies as in figure 4 the circle is deformed. This indicates that at middle high frequencies where the ILD cues are more prominent, the EI patterns indicate a dependency on the elevated sources which could be explained by the fact that short-wavelength sounds are not diffracted around the head to the same extent as long wavelengths.

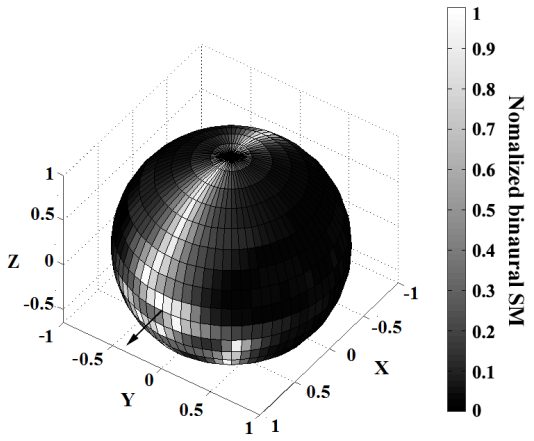

Figure 4: The results of the comparison of the EI patterns at $f \approx 9.4$ kHz for a sound source at $\hat{\phi} = \hat{\theta} = 0^\circ$ by using the vertical-polar coordinate system. The colour bar indicates the value of eq. 5 normalised by its maximum value.

3.2. Monaural cues

One of the main characteristics in the analysis of the head related transfer functions (HRTFs) is the spectral colouration introduced by the outer ear. Prominent peaks and notches can be found at different frequency ranges that are considered potential cues for elevation. For instance, the ambiguity on a cone of confusion can be discriminated with the appropriate spectral cues that reside mainly at 8 - 16 kHz [20], while for up-down location the appropriate spectral cues reside mainly at 6 - 12 kHz [20].

Additionally, it has been shown that the tonotopic organization in the cochlea is preserved in the higher regions of the auditory path such as in the cochlea nucleus (CN) [21]. As a consequence, it was considered necessary to check whether the peaks and notches of the HRTFs could be preserved in the TA patterns (eq. 4).

In Figures 5, 6 we can see[1] from visual observations that all the pinna resonances and pinna nulls of the HRTFs are preserved in the TA patterns but in a rather smoothed out representation. This smooth representation of the TA patterns is due to the lower frequency resolution of the channels of the gammatone filterbank (100 frequency channels) compared to the finest resolution of the HRTFs and the compressive character of the square root compressor in the peripheral processing unit which changes the dynamic range of the signal.

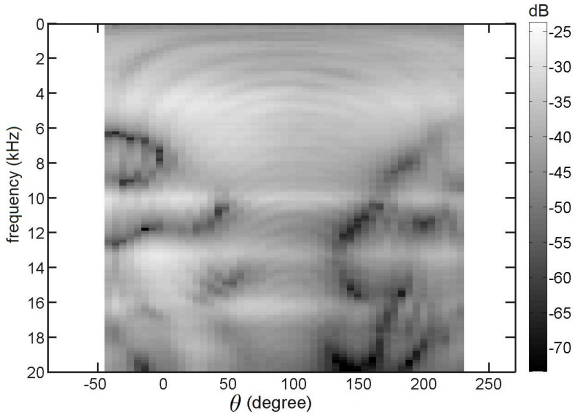

Figure 5: The HRTF of a KEMAR with a small pinna from the CIPIC database (subject 165, right ear) [16] in the median plane ($\phi = 0^\circ$) in the interaural-polar coordinate system.

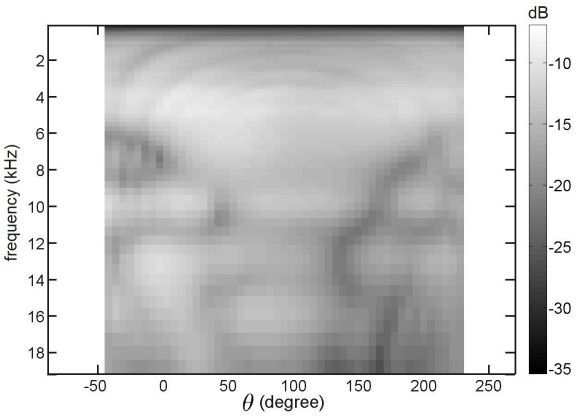

Figure 6: The TA patterns as they have been created by the HRTF of a KEMAR with a small pinna from the CIPIC database (subject 165, right ear) [16] in the median plane ($\phi = 0^\circ$) in the interaural-polar coordinate system.

[1] Although the results depict the monaural processor produced by the right ear, similar results could also be found at the corresponding TA patterns of the left ear.

3.3. Decision making device

Considering that the SM of the monaural and binaural cues have been combined as indicated in eq. 11 it was considered necessary to further explore the influence of the γ_{bin} (eq. 6) and σ_{mon} (eq. 9) parameters in the final stage of the model independently. Figures 7 - 10 illustrate the effect of the γ_{bin} and σ_{mon} in the binaural and monaural SMs for very high and very low values at a position exactly in front of a KEMAR ($\hat{\phi} = \hat{\theta} = 0^{\circ}$), for a sound source as described in sec. 4. For the binaural SM, eq. 6, (Figures 7, 8), which is responsible for giving the highest similarity to all the points around the target azimuth angle independent of the elevation angle, it can be noticed that the γ_{bin} parameter spreads the values around the target azimuth angle $\hat{\phi} = 0^{\circ}$. This implies that the binaural SM is roughly independent of the elevation angle ($s_{bin}(\phi, \theta) \approx s_{bin}(\phi)$) which indicates the lack of EI cues to match to all the EI patterns along the median plane.

In contrast, the monaural SM shows a different behavior. For high values of σ_{mon} (Figure 9), the TA cues around the median plane match with all the TA patterns indicating in this way a high chance the sound source is located at a position outside that region. Nevertheless at all locations the SM has a rather low value which ranges from 0.85-1.0. In cases where σ_{mon} is less than one (Figure 10) the performance of the monaural processor improves, and for extremely low values, the monaural processor gives the highest similarity at the point where a sound source is located.

Based on the behavior of the γ_{bin} parameter and the fact that the EI patterns are associated with the ITD and ILD cues, we could conclude that the binaural SM ($s_{bin}(\phi, \theta)$) is able to give an estimation of the position of the sagittal plane with the γ_{bin} parameter restricting or expanding the predicted region around the estimated sagittal plane. In addition, considering that TA patterns are associated with the spectral cues, the monaural SM is able to predict the exact location of a sound source with the σ_{mon} parameter restricting or expanding the predicted region around the estimated location. However, this is not only limited to the target position but it expands to other locations as well, where the TA patterns are quite similar. This is associated with the lack of the spectral cues to resolve the exact location of a sound source on a cone of confusion as indicated in Figure 10 where there is a high probability for a sound source at $\hat{\theta} = 180^{\circ}$.

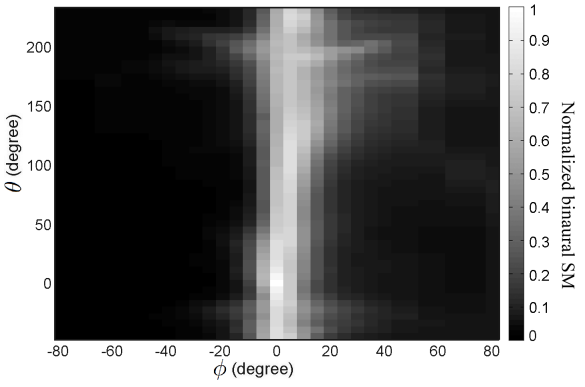

Figure 7: The prediction of the binaural pattern matching process (eq. 7) normalized by its maximum value for a white noise sound source as described in [22] at $\hat{\phi} = 0^{\circ}$ and $\hat{\theta} = 0^{\circ}$ in the interaural-polar coordinate system and for a high value of the γ_{bin} parameter ($\gamma_{bin} \gg 1$).

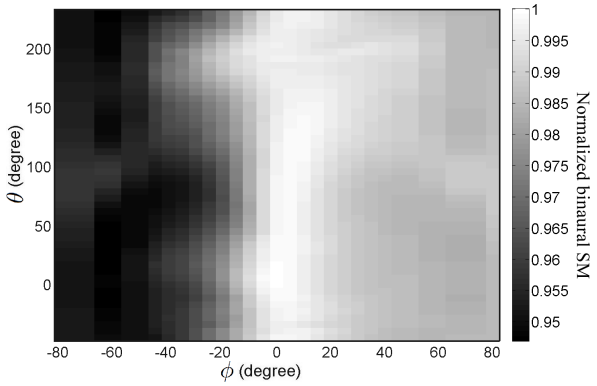

Figure 8: The prediction of the binaural pattern matching process (eq. 7) normalized by its maximum value for a white noise sound source as described in [22] at $\hat{\phi} = 0^{\circ}$ and $\hat{\theta} = 0^{\circ}$ in the interaural-polar coordinate system and for a low value of the γ_{bin} parameter ($\gamma_{bin} \ll 1$).

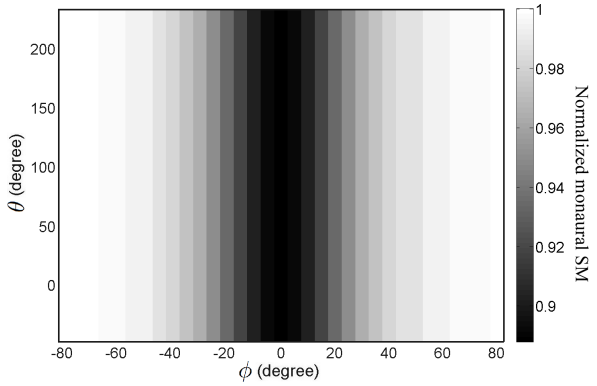

Figure 9: The prediction of the monaural pattern matching process (eq. 10) normalized by its maximum value for a white noise as a sound source at $\hat{\phi} = 0^{\circ}$ and $\hat{\theta} = 0^{\circ}$ in the interaural-polar coordinate system and for a high value of the σ_{mon} parameter ($\sigma_{mon} \gg 1$).

4. COMPARISON TO LISTENING TESTS

In order to validate the performance of the proposed model, the experimental data of Makous and Middlebrooks [22] have been used. In the particular listening test six listeners with normal hearing had to identify the actual location of a sound source at different locations in 3D space at a fixed distance of 1.2m in an acoustic environment with 40 dB SPL ambient noise and a room that can be considered anechoic for frequencies above 500 Hz. The sound source had a sound pressure level that ranged randomly for each trial from 40 to 50 dB sensation level and a frequency range between 1.8 kHz and 16 kHz. From the two experiments that were conducted we are mainly interested in the so called open-loop trials, in which the duration of the stimulus was 150ms and the subject had his/her head at a fixed position. In this way any dynamic cues that could have been created were excluded. Finally across all subjects, each stimulus location was tested in total 31 times giving an azimuth and elevation mean error and standard deviation for each subject.

Figures 11 - 13 illustrate the prediction of the model for a vir-

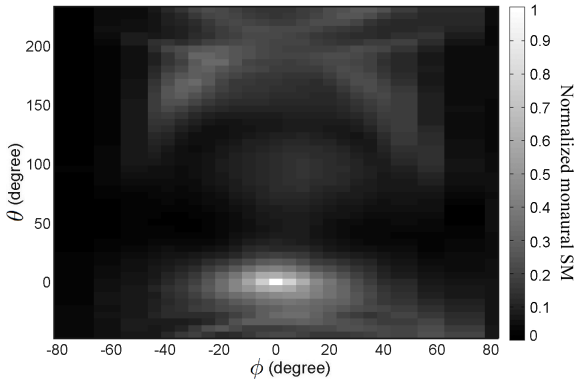

Figure 10: The prediction of the monaural pattern matching process (eq. 10) normalized by its maximum value for a white noise as a sound source at $\hat{\phi} = 0^\circ$ and $\hat{\theta} = 0^\circ$ in the interaural-polar coordinate system and for a low value of the σ_{mon} parameter ($\sigma_{mon} \ll 1$).

tual sound source[2] with the same specifications of the listening test[3] at three different positions. The center of the ellipses on the Figures indicate the average error of the detected sound source position in the listening tests and it has been calculated by averaging the mean error of the response of each subject. The average error is characterized by its mean value, which is the center of the ellipses, and the standard deviation about the mean value, which is not indicated. The length of transverse and conjugate diameters indicate the average standard deviation about the mean response for each subject for the azimuth and elevation angle correspondingly and it has been calculated by averaging the standard deviation of the response of each subject. The average standard deviation of the azimuth and the elevation angle is characterized by it's mean value, which is the length of transverse and conjugate diameters correspondingly, and a standard deviation about this value, which is not indicated. The parameters γ_{bin} and σ_{mon} of the model have been adjusted in such a way to fit as closely as possible to the listening test results, where $\gamma_{bin} = 1.82$ and $\sigma_{mon} = 0.3$.

Although the performance of the model, from visual observation of the figures 11 - 13, seem to give quite a good prediction of the results of the listening tests, some other aspects should be considered. Due to the fact that the frequency range of the sound source is between 1.8 kHz and 16 kHz all the information that is hidden in the low frequencies for the ITDs has been eliminated. This results in the total SM being spread along the estimated sagittal plane, something that is influenced by the fact that the EI patterns are only using the ILDs and the envelope of the ITD cues.

Despite the fact that the average error and the average standard deviation of the detected sound source position have been used for the creation of the ellipses of the listening tests, the actual errors are even higher. For instance for a sound source in the median plane at an elevated position at $\hat{\theta} = 45^\circ$ (Figure 11), the average error can vary from $2.7^\circ \pm 4.1^\circ$ for the horizontal dimension[4] and $-5.9^\circ \pm 10.6^\circ$ for the vertical dimension while the average standard deviation can vary from $3.0^\circ \pm 2.3^\circ$ for the horizontal dimension and $7.9^\circ \pm 2.0^\circ$ for the vertical dimension. This means that in general the error of the estimated location of the horizontal dimension can vary from -2.1° to 12.1° and from 18.6° to 59.6° for the vertical dimension. Furthermore, these estimated values do not consider the front-back confusion errors, something that is depicted by the prediction of the model.

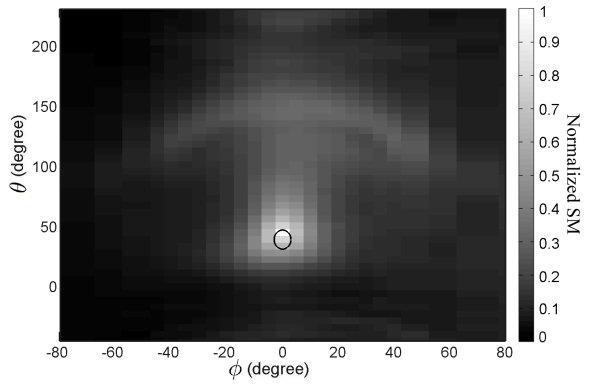

Figure 11: The prediction of the perceptual model (eq. 11) normalized by its maximum value for a sound source at $\hat{\phi} = 0^\circ$ and $\hat{\theta} = 45^\circ$ in the interaural-polar coordinate system and the listening test results (ellipse) of Makous and Middlebrooks [22].

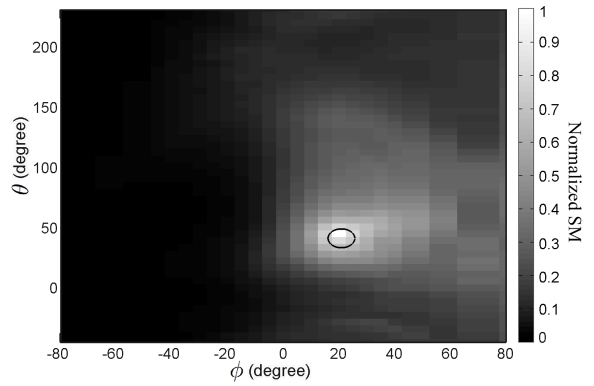

Figure 12: The prediction of the perceptual model (eq. 11) normalized by its maximum value for a sound source at $\hat{\phi} = 20^\circ$ and $\hat{\theta} = 45^\circ$ in the interaural-polar coordinate system and the listening test results (ellipse) of Makous and Middlebrooks [22].

5. CONCLUSIONS

The aim of the current study was to explore some of the characteristics of a biologically inspired model and to illustrate its performance in comparison to real listening tests. The results of the listening test indicate that the current model is able to predict, at least qualitatively, the human performance in localization tests of stationary sounds. Nevertheless, further investigation is necessary for a quantitative analysis of the model and a better quantification of the range that γ_{bin} and σ_{mon} should vary to predict the human performance in the localization of broadband sound sources with individualized or generalized HRTFs.

[2]The HRTFs that have been used are from the CIPIC database[16].

[3]The sound pressure level has been considered to be on average 45 dB SPL.

[4]In the notation $m \pm \sigma$ the first value indicates the mean value, while the second the standard deviation around this value.

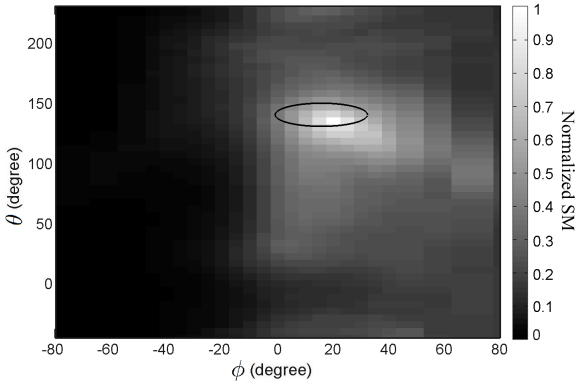

Figure 13: The prediction of the perceptual model (eq. 11) normalized by its maximum value for a sound source at $\hat{\phi} = 20°$ and $\hat{\theta} = 135°$ in the interaural-polar coordinate system and the listening test results (ellipse) of Makous and Middlebrooks [22].

6. ACKNOWLEDGMENTS

The research for this paper was financially supported by Meridian Audio Ltd. and the University of Southampton. In developing the ideas presented here, I have received helpful input from Dr. Stephan Bleeck from the University of Southampton, and Prof. Ville Pullki from the Aalto Department of Signal Processing and Acoustics. Very many thanks also to Dr. T. Takeuchi, Dr. M. Park and Prof. J. C. Middlebrooks , amongst others, for useful feedback and advice.

7. REFERENCES

[1] J Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization*, MIT, Cambridge, MA, 1997.

[2] Lloyd A Jeffress, "A place theory of sound localization," *Journal of Comparative and Physiological Psychology*, vol. 41, no. 1, pp. 35–39, 1948.

[3] N. I. Durlach, "Equalization and Cancellation Theory of Binaural Masking-Level Differences," *The Journal of the Acoustical Society of America*, vol. 35, no. 8, pp. 1206–1218, 1963.

[4] G. Von Bekesy, "Zur Theorie des Hörens: Über das Richtungshören bei einer Zeitdifferenz oder Lautstärkeungleichheit der beidseitigen Schalleinwirkungen," *Phys Z*, pp. 824–838, 1930.

[5] Erno H A Langendijk and Adelbert W Bronkhorst, "Contribution of spectral cues to human sound localization," *The Journal of the Acoustical Society of America*, vol. 112, no. 4, pp. 1583–1596, 2002.

[6] Robert Baumgartner, Piotr Majdak, and Bernhard Laback, "Assessment of sagittal-plane sound-localization performance in spatial-audio applications," in *The technology of binaural listening*, pp. 93–120. 2013.

[7] Paul M Hofman and A John Van Opstal, *Spectro-temporal factors in two-dimensional human sound localization*, vol. 103, ASA, 1998.

[8] Joyce Vliegen and A John Van Opstal, "The influence of duration and level on human sound localization.," *Journal of the Acoustical Society of America*, vol. 115, no. 4, pp. 1705–1713, 2004.

[9] Munhum Park, Philip A. Neslon, and Kyeongok Kang, "A Model of Sound Localisation Applied to the Evaluation of Systems for Stereophony," *Acta Acustic United with Acustica*, vol. 94, pp. 825–839, 2008.

[10] R.D. Patterson, K. Robinson, J. Holdsworth, D. McKeown, C. Zhang, and M. Allerhand, "Complex sounds and auditory images," in *Auditory physiology and perception, 9th International Symposium on Hearing*, Y. Cazals, L. Demany, and K. Horner, Eds., Oxford, 1992, pp. 429–446, Pergamon, 1992.

[11] Donald D Greenwood, "What is "Synchrony suppression"?," *The Journal of the Acoustical Society of America*, vol. 79, no. 6, pp. 1857–1872, 1986.

[12] R C Kidd and T F Weiss, "Mechanisms that degrade timing information in the cochlea.," *Hearing Research*, vol. 49, no. 1-3, pp. 181–207, 1990.

[13] T Dau, D Püschel, and A Kohlrausch, "A quantitative model of the "effective" signal processing in the auditory system. I. Model structure.," *Journal of the Acoustical Society of America*, vol. 99, no. 6, pp. 3615–3622, 1996.

[14] Sergios Theodoridis and Konstantinos Koutroumbas, *Pattern Recognition*, Academic Press, 4th edition, 2009.

[15] Symeon Mattes, Philip Arthur Nelson, Filippo Maria Fazi, and Michael Capp, "Towards a human perceptual model for 3D sound localization," in *28th Conference on Reproduced Sound: Auralisation: Designing With Sound*, 2012.

[16] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano, "The CIPIC HRTF database," in *Proc. 2001 IEEE Workshop on Applications of Signal Processing to Audio and Electroacoustics*, pp. 99–102, 2001.

[17] Richard M Stern, Andrew S Zeiberg, and Constantine Trahiotis, "Lateralization of complex binaural stimuli: A weighted-image model," *The Journal of the Acoustical Society of America*, vol. 84, no. 1, pp. 156–165, 1988.

[18] Ewan A Macpherson and Andrew T Sabin, "Binaural weighting of monaural spectral cues for sound localization.," *Journal of the Acoustical Society of America*, vol. 121, no. 6, pp. 3677–3688, 2007.

[19] Frederic L Wightman and Doris J Kistler, "Monaural sound localization revisited," *The Journal of the Acoustical Society of America*, vol. 101, no. 2, pp. 1050–1063, 1997.

[20] Henrik Møller and Daniela Toledo, "The Role of Spectral Features in Sound Localization," *Audio Engineering Society Convention 124/7450*, 2008.

[21] R E Wickesberg and D Oertel, "Tonotopic projection from the dorsal to the anteroventral cochlear nucleus of mice.," *Journal of Comparative Neurology*, vol. 268, no. 3, pp. 389–399, 1988.

[22] James C Makous and John C Middlebrooks, "Two-dimensional sound localization by human listeners," *The Journal of the Acoustical Society of America*, vol. 87, no. 5, pp. 2188–2200, 1990.